SABS STATISTICS PRACTICAL

DAVID STEINSALTZ

The first 3 questions are intended for the first practical session.

(1) Suppose the mean household income in a city is £30,000, and the standard deviation is £15,000. 1000 households are selected at random.
  (a) Estimate the probability that the average household income in the sample is no more than £31,000. Does it matter how many households there are in the city? Does it matter what the distribution of incomes is?
  (b) Suppose we do not know the true mean for the whole city, but the mean of our sample is £30,000. Calculate 95% and 99% confidence intervals for the mean of the whole population.

(2) Suppose we have a new IQ test, which we wish to compare to a standard IQ test, to see whether the scores on the new test have the same average as the scores on the old test. We take 200 subjects, and pick 120 at random to take the new test; the other 80 take the old test. The results are as follows:

Test	# subjects	Mean	SD
Old	80	101	11
New	120	98	8

We wish to do a Z test at the 0.01 level to determine whether the means on the tests could be the same.
  (a) State the test hypotheses.
  (b) State any assumptions you have to make.
  (c) Calculate the test statistic, compute the p-value.
  (d) Compare to the critical value and conclude.

(3) Suppose you have 4 independent observations from a normal distribution with unknown mean and unknown variance. You estimate the mean and SD from the data, and use them to compute the Student t statistic $T = (\bar{X} - 1)/(\hat{\sigma}/\sqrt{n})$. You report it to someone, who mistakenly believes that the variance is known, and thus performs a Z test for the hypothesis that the mean of the distribution is 1. What will the impact be on the statistical conclusions? What if the error goes in the other direction (that is, you compute from a known variance, but someone believes you have delivered a T statistic)? Carry out some simulations to demonstrate the error that will be made.
  (a) If the person is performing two-tailed hypothesis tests at level 0.05, calculate the size of the error in each case.
  (b) Carry out some simulations to demonstrate this.

(4) A study was carried out to test the prevalence of side effects from the pertussis vaccine.[1] Of 339 infants who received their first injection of vaccine, 69 showed adverse reactions.
  (a) Compute 95% and 99% confidence intervals for the probability of an adverse reaction to the vaccine.

Date: 6 December, 2019.

[1] Published as "Whooping-cough vaccination: An assessment," Miller et al., The Lancet, 1974. Described in Samuels and Witmer, *Statistics for Life Sciences*, p.212.

(b) What do these confidence intervals mean? What assumptions go into the interpretation?

(5) The table below shows measurements of height and forced expiratory volume in one second (FEV) in a sample of male medical students.

Height (cm)	FEV (litres)
164.0	3.5
170.4	3.2
171.3	3.2
172.0	3.8
176.0	3.8
177.0	5.4
178.0	3.0
181.0	4.0
183.7	4.7

(a) Represent the data graphically.
(b) Calculate the sample correlation.
(c) Perform a linear regression and include the fitted line in your graph.
(d) Test whether the slope of the regression line differs significantly from zero.

(6) To determine the effectiveness of a new drug on the level of haemoglobin in the blood of anemic patients, 10 randomly selected patients who underwent this treatment were sampled. The table below shows the level of haemoglobin in the patients' blood before and after the treatment.

Patient	1	2	3	4	5	6	7	8	9	10
Before	11.2	9.4	9.9	9.3	8.9	8.2	10.5	8.8	10.3	9.8
After	12.9	10.8	10.3	10.9	8.5	8.9	10.4	8.5	11.2	10.1

Carry out both the two-sample t test and the paired sample t test at a significance level of 0.05 to determine if the drug is effective in increasing the level of haemoglobin in the blood. Given the setting of the experiment, which test is more appropriate? Why?

(7) You have a small number (2, 3, 5, or 10) of independent samples from an unknown distribution with expectation 1. You perform a T test at level 0.01 for the null hypothesis that the expectation is 1. However, the distribution that you are sampling from is not normal.
(a) Simulate this experiment 1000 times with each of a number of different distributions: Poisson with parameter 1, Exponential, Uniform on [0, 2]. Under which circumstances does the test work (in the sense that the probability of rejecting a true null hypothesis is close to 0.01)?
(b) Compare the distribution of your simulated statistic to Student t distribution using a Q-Q plot. (Note: The command `qqplot` makes a Q-Q plot for comparing two different data sets. One of them can be a large sample from the correct T distribution.)
(c) Try to formulate a guess about which properties of a distribution are required for the t test to be accurate, or for it to be conservative.

(8) The following table gives the observed counts in 1-second intervals of alpha particles emitted from a radioactive source. Use a χ^2 test at the 5% level to assess the goodness of fit to the Poisson distribution.

n	Observed
0	5267
1	4436
2	1800
3	534
4	111
5+	21

(9) ***E. coli* descriptive statistics in `R`** (Optional, for morning or afternoon.) In the following sequence of questions will be using data from the completely sequenced genome of *E. coli*, more precisely strain K-12, substrain MG 1655, version M52 short. *Adapted from a lab written by Terry Speed and Bin Yu.*

(a) **Background and raw data.**
The sequence data have come from the NCBI website: `www.ncbi.nlm.nih.gov.` You may download the data by typing
`load(url('http://steinsaltz.me.uk/DTC/ecoli.rda')).`
What was originally a single string of length 4641652 has been processed by the command `strsplit(x,'')` to create a vector `ecp` of length 4641652, each entry consisting of one letter of A, C, G, T.

(b) Count the number of Adenines, Guanines, Cytosines and Thymines (As, Gs, Cs and Ts). Then there is the dinucleotide composition, that is, the number of AAs, AGs etc. (along one strand). There further is the trinucleotide composition, that is, the number of AAAs, AAGs etc.

(i) If you had to assign a probability to observing an A at a stated position on the *E. coli* genome, what figure would you use and why? Under what assumptions does this seem appropriate?

(ii) Type `L=length(ecp)` followed by `table(ecp[-L],ecp[-1])`. What is this telling you?

(iii) Your are told there is an A at a position along the *E. coli* genome. What probability would you assign to it being followed by a G, and why?

(iv) Does the *E. coli* composition data suggest that the event we observe a G at one site is independent (in some suitable sense) of the previous two bases? Explain fully, illustrating with appropriate data.

(c) **Purine counts.**
Divide up the data into about 46,400 blocks of 100 base pairs (bp). Count the number of purines (i.e. A or G). Do the same for the 4,640 blocks of 1,000 bp, and 464 blocks of 10,000 bp.

(i) For each set of counts, calculate the mean and standard deviation of the number of purines per block, and draw histograms of these numbers.

(ii) Compare the results of (a) across the different block sizes and comment.

(iii) For each block size, calculate the fraction of the counts within 1, 2 and 3 standard deviations of the mean.

(iv) Repeat (a), (b) and (c) for *proportions* (rather than counts) of purines in each block.

(d) Compute counts of TATAAT in blocks of 5,000 bp. Assuming that these counts follow a Poisson distribution, estimate the parameter of this distribution and obtain an estimate of the standard error of your parameter estimate. This can be done by either a formula or by simulation.